## SUPPLEMENTARY MATERIAL

# 1 PDBEST: GENERAL FILTERING PROTOCOL

We have elaborated a general PDB filtering protocol using PDBest, that encompass the most important and general pre-processing tasks that might help users to build a curated structural database. The main contemplated tasks on this tutorial are described below:

- Select structures based on experimental method (e.g, X-ray crystallography only);
- Identify non-protein chains;
- Remove identical chains (similar sequences at 100%);
- Select structures based on X-ray resolution and r-value;
- Split files by chain;
- Filter models;
- Identify chains containing missing atoms or residues;
- Filter atom with multiple occupancies (keep only the greater one);
- Convert between file formats (e.g, PDB, mmCIF, FASTA);
- Remove water molecules;
- Remove ANISOU record type;

The general protocol file is available at:
http://www.pdbest.dcc.ufmg.br/protocol/general-protocol.ses

# 2 CASE STUDIES

In this section we describe successfully developed studies by our group and co-workers where the use of PDBest was key.

## 2.1 Case #1: Studying contacts in globular proteins

Studies about how to reliably prospect inter-residue contacts in proteins using diferent methodologies have been performed by our group [1] and co-works. In this work the authors conducted a comparative analysis between two classical approachs: the traditional cutoff dependent (CD) and cutoff free Delaunay tessellation (DT). A database was built, comprising three main protein structural classes: all alpha, all beta, and alpha/beta. The following general selection criteria was performed with PDBest:

1. **Online Query:** Select on the SCOP Classification Browser the classes *all alpha* OR *all beta* OR *alpha/beta* AND X-ray resolution $\leq$ 2.0Å AND Refinement R Factor *R-Work* $\leq$ *0.2*. Also remove similar sequences with a 30% similarity threshold. Only select chains containing from 50 to 600 amino acid residues.
2. **Download and verify inconsistences:** After using the query presented in the previous step, the PDB files downloaded were examined. Files with potentially harmful inconsistencies which could introduce biases to the contact analysis were discarded. These included gaps on the structure and missing atoms.
3. **Pre-Process:**
    - Re-enumerate residues and atoms;
    - Exclude chains with the same sequence (as given by the SEQRES record);
    - Delete chains with missing atoms;
    - Discard chains with non-standard amino acids (with the exception of selenomethionine);
    - Only keep atoms with greatest occupancy on Coordenate Section;
    - Keep only the first model on Coordenate Section.

Protocol file available at:
http://www.pdbest.dcc.ufmg.br/protocol/glogin-protocol.ses

### 2.2 Case #2: Studying Cross-Inhibition in Serine Proteases

In a previous work of cross-inhibition [2], the PDBest plataform was used successfully to query the Protein Data Bank and pre-process PDB files of complexes between Serine Proteases and the inhibitor Eglin C. The goal of the work was to study structural patterns on the contact interface between the protein and its ligands and understand why different proteins are inhibited by the same proteic ligand, Eglin C. Several steps were used to acquire and treat a set of target molecules in the study. The steps are described below:

1. **Online Query:** Query structures with macromolecule name *"Eglin C"* and two assimetric chains, with X-ray resolution $\leq$ 2Å.
2. **Download and verify inconsistences:** The PDB files were downloaded and examined to discover possible annotation or submission errors. PDBest provides a report indicating missing atoms and residues apart from the identification of multiple occupancies and non-standard residues. It is possible to decide to keep or discard the files based on this report.
3. **Pre-Process:** The PDB files were filtered considering the following set of requirements:

   - Delete hydrogens;
   - Split files by chain;
   - Keep atoms and residues original numbering on the Coordenate Section;
   - If multiple occupancies are identified, keep only the greater one;
   - Remove solvent molecules, HETATM and anisotropy records on the Coordenate Section.

Protocol file available at:
http://www.pdbest.dcc.ufmg.br/protocol/cross-inhibition-protocol.ses

### 2.3 Case #3: Molecular Recognition on Protein Kinases (CDK)

Noncovalent interactions are driving forces guiding the molecular recognition between proteins and ligands. Particularly, for the CDK (Cyclin-Dependent Kinase) protein family this phenomenon have attracted great interest due to binding site promiscuity. In other words, a vast number of diferent ligands bind the same binding site. In a work developed over a Summer Course at Universidade Federal de Minas Gerais-Brazil[1] this phenomenon was analised through a visual approach and PDBest was essential to query and pre-process a set of complexes deposited in the PDB.

1. **Online Query:** Perform a text search using *"CDK2 in complex with inhibitor RC"*.
2. **Download and verify inconsistences:** The PDB files were downloaded and examined to discovery some possible annotation or submission errors. The PDBest provides a report indicating about missing atoms and residues besides to identify occupancy and no-standard residues. In this case, all files presenting any missing residues were discarded and multiple occupancies repaired as described in the next step.
3. **Pre-Process:**

   - Delete hydrogens;
   - Split files by chain;
   - Keep atoms and residues original numbering on the Coordenate Section;
   - If multiple occupancies are identified, keep only the greater one;
   - Remove solvent molecules and anisotropy records on the Coordenate Section.

Protocol file available at:
http://www.pdbest.dcc.ufmg.br/protocol/protein-kinase-protocol.ses

## REFERENCES

[1]C H da Silveira, D E V Pires, R C Melo-Minardi, C Ribeiro, C J M Veloso, J C D Lopes, W Meira Jr, G Neshich, C H I Ramos, R Habesch, and M M Santoro. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function and Bioinformatics*, 74(3):727–743, 2009.

[2]V M Gonçalves-Almeida, Douglas E V Pires, Raquel Cardoso Melo-Minardi, Carlos Henrique da Silveira, W Meira Jr., and Marcelo M Santoro. HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342–349, 2012.

[1] http://www.lbs.dcc.ufmg.br/cvbioinfo/